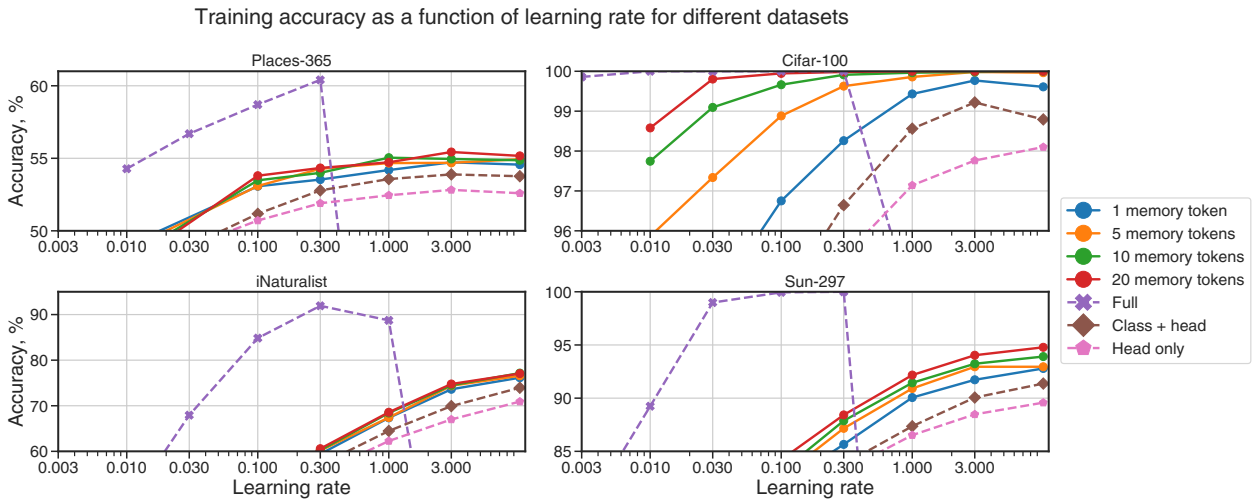

Figure 1. Performance of memory-augmented transformers for different datasets and comparison with baseline fine-tuning methods.

A. Dependence of training accuracy on number of memory tokens

As we discussed in the experiments sections, the test data performance as a function of number of memory tokens appear not to improve much. We hypothesize that the reason for that might be the growing generalization gap. Specifically on Fig. 1 we show the training performance of memory-augmented transformers. As can be clearly seen, there the performance improves monotonically as the number of memory token grow.

B. Exploring memory attention

In this section we explore how the attention to memory changes as training progresses. In Fig. 2 we show how the attention changes over training trajectory. Specifically, we measure the fraction of input samples in the validation subset, that have at least one input token, have cumulative attention to memory of at least 0.5 at different layers. We used the first 3 heads and measured the attention at 4 different layers in the beginning, in the middle and at the top of the network. As we can see, the general pattern is that attention to memory tends to increase as learning progresses.

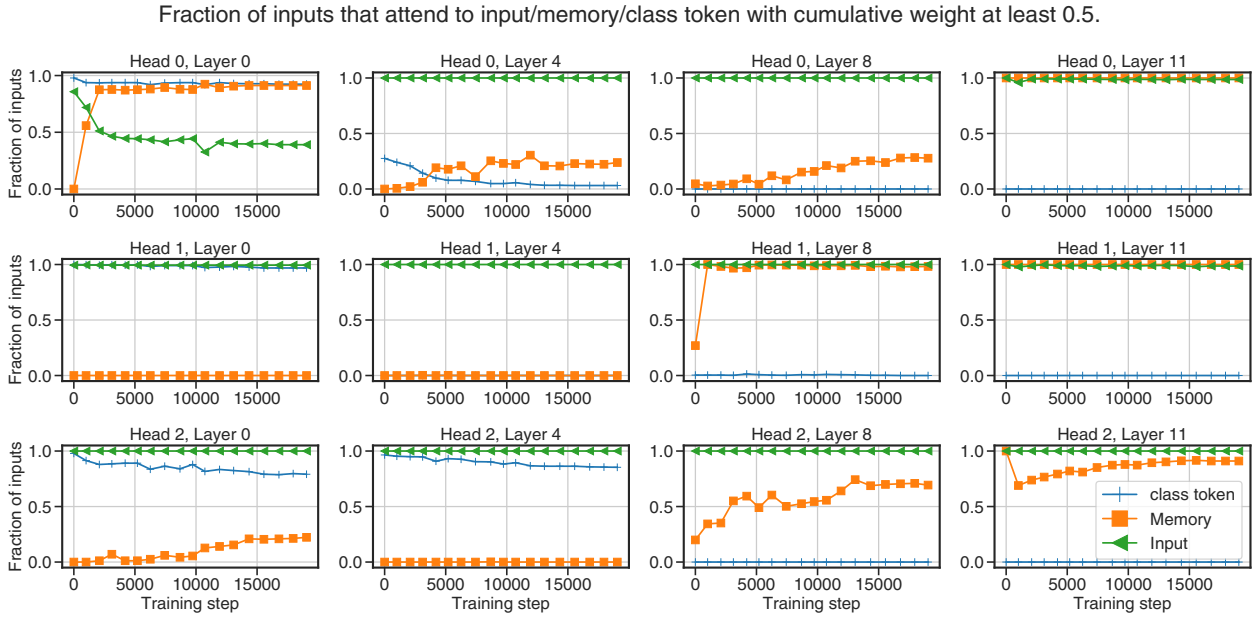

Figure 2. Attention of input tokens to different types of other tokens for individual heads and layers. We only include 3 heads (0, 1, and 2) and 4 layers spread uniformly over architecture. Here we calculate what fraction of samples have at least one token, that attends with weight at least 0.5 to (a) memory (■) (b) class token (+) and (c) self-attention (◀). Remarkably, we see a significant variability for different heads.